

《庚巳編》舊版造字轉碼說明

中研院資訊所文獻處理實驗室
中研院史語所漢籍電子文獻工作小組
2008/8/5 葉冠孜 製作

一、《庚巳編》一書(庚巳編.xml)使用舊版造字 108 個，字頻 515 次，詳如附件一。這 108 個造字中，85 個可轉成 Windows XP 能顯示的字，字頻 412 次；另外 23 個字必須轉成構字式，字頻 103 次。

二、附件一的造字分析表說明如下：

甲、編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

乙、造字：舊版造字

丙、字頻(txt)：造字在「.txt」文件的出現次數

丁、字頻(xml)：造字在「.xml」文件的出現次數

戊、Big5：造字的 Big5 碼

己、Unicode：造字所對應的 Unicode 碼

庚、WinXP：造字在 Windows XP 的對應字形

辛、構字式：Windows XP 若無對應字形，則改採用構字式

壬、備註凡例：

- 1、校對問題，新版漢籍錯字，人工取代：新版漢籍因字句缺漏，而造成頻次不同，處理方式為將缺漏的字句補上，並將沒被修改到的舊漢籍造字，一併以對應的 Unicode 或構字式取代。如編號 1544 的「絙」字，在 xml 檔少了一個頻次，經比對原書發現在「<qe>絙夢身中鄉試六十七名原本闕「六」、「七」二字，據說庫本補。</qe>」句中，xml 檔少了「絙」字，因此將「絙」字補上，其他未經修改的部分，則直接手動取代為「絙」字。
- 2、異體字問題：新版漢籍考量到使用者檢索及使用時的便利性，將用字原則改為除專詞等特殊情形之外，一律改用標準字呈現。如編號 4132 的「血𠂇」係「眾」字之異體，故以「眾」字取代。
- 3、Unicode 字型呈現差異：Unicode 字型與舊漢籍造字有些微差異，但只是字體風格差異，實際上仍為同一個字，因此仍取 Unicode 字型。如編號 3846 的「壻」字，Unicode 字型呈現為「壻」，實際上仍為同一字。

三、Unicode 目前收錄的漢字總數為 70194，分屬於三個不同區段，詳如表一。目前 Windows XP 只支援 CJK 認同表意文字區的 20902 個字，內碼為 4E00-9FFF。所以造字編號 4808 的「健」字，Unicode 編碼為 3493，由於 Windows XP 並不支援，仍須使用構字式「亻厶達」。

表一、Unicode 的字數及編碼區段

Unicode	字集子集合	新增字數	新增編碼區段	總字數	WinXP
1.1 版	CJK 認同表意文字區	20902	4E00-9FFF	20902	支援
3.0 版	CJK 認同表意文字擴充 A 區	6582	3400-4DFF	27484	不支援
3.1 版	CJK 認同表意文字擴充 B 區	42710	20000-2A6D6	70194	不支援

四、txt 檔與 xml 檔頻次不同且造字分析表「備註」欄有備註問題情況者，詳細取代內容可參考附件二的手動取代表。附件二的手動取代表說明如下：

- 甲、編號：Big5 造字空間為 6217 個，編號由 1 到 6217。
- 乙、造字：舊漢籍造字。
- 丙、檔案原文摘錄：摘錄檔案(庚巳編.xml)中包含該造字的文句段落，其中紅字底線的部分，為需要手動取代的原文。文句後方標示(全部取代)者，表示執行全部取代的動作，不一一列舉所有原文。
- 丁、手動取代結果：變更過後的檔案內容，其中藍字底線的部分，為修改後的結果。文句後方標示(全部取代)者，表示執行全部取代的動作，不一一列舉所有修改後結果。


附件一、《庚巳編》造字分析表

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
38		8	8	FA65			王厶已	玓，異體字問題。
44	颶	1	1	FA6B	295D7		風厶貝	
296	厶	11	11	FBEC			厶	

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
297	𠂔	9	9	FBED			𠂔	
299	𠂔	2	2	FBEF			𠂔	
300	𠂔	2	2	FBF0			𠂔	
346	电	1	1	FC5F			电	
452	火	1	1	FCEB			火	
717	足	1	1	FEBA			足	
738	心	1	1	FECF			心	
770	扌	3	3	FEEF			扌	
994	楚	4	4	8F73	8E75	楚		
1000		9	9	8F79			王𠂔巳	玨，異體字問題。
1010	珏	1	1	8FA5	73CF	珏		
1400	窓	5	5	91F1	7A93	窓		
1544	絙	9	8	92E4	7D59	絙		校對問題，新版漢籍 錯字，人工取代。
1595	匜	1	1	9358	20937		匚△𠂔	
1618	𠂔	1	1	936F	2522D		目𠂔𠂔	
1642	脉	3	3	93A9	8109	脉		

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
1691	鬚	1	1	93DA	4BFC		髟彡狄	
1926	覓	1	1	9569	8994	覓		
1981	讐	2	2	95C2	8B90	讐		
1991	鯨	1	1	95CC	9C1F	鯨		
1999	贊	1	1	95D4	8CDB	贊		
2044	踈	3	3	9642	8E08	踈		
2143	火	3	3	96C7			火	
2610	鱣	5	5	99C3	9C53	鱣		
2824	鯁	1	1	9AFC	9BFE	鯁		
3184	櫛	2	2	9D6B	6AC9	櫛		
3255	善	43	43	9DD4	5584	善		
3789	冲	12	12	8154	51B2	冲		
3805	勅	2	2	8164	52C5	勅		
3814	却	18	18	816D	5374	却		
3817	廁	2	2	8170	53A0	廁		
3819	廝	10	10	8172	53AE	廝		

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
3830	咤	2	2	817D	54A4	咤		
3834	嗶	1	1	81A3	5637	嗶		
3846	壻	1	1	81AF	58FB	壻		Unicode 呈現差異。
3850	姪	7	7	81B3	59D9	姪		
3858	寃	5	5	81BB	5BC3	寃		
3864	峯	7	7	81C1	5CEF	峯		
3868	卮	1	1	81C5	5DF5	卮		
3883	徧	7	7	81D4	5FA7	徧		
3889	慤	1	1	81DA	6164	慤		
3891	慤	1	1	81DC	6159	慤		
3932	稜	5	5	8246	6936	稜		
3935	槩	4	4	8249	69E9	槩		Unicode 呈現差異。
3966	烟	5	5	8268	70DF	烟		
3974	牀	6	6	8270	7240	牀		
3978	猪	1	1	8274	732A	猪		Unicode 呈現差異。
3979	猫	6	6	8275	732B	猫		

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
3996	疎	1	1	82A8	758E	疎		
4012	鼓	3	3	82B8	76B7	鼓		
4013	盃	2	2	82B9	76CC	盃		
4016	着	11	11	82BC	7740	着		Unicode 呈現差異。
4024	礮	2	2	82C4	792E	礮		
4039	竝	1	1	82D3	7ADA	竝		
4040	竝	1	1	82D4	7ADD	竝		
4041	竝	1	1	82D5	7AEA	竝		
4044	筋	1	1	82D8	7B6F	筋		
4069	綑	1	1	82F1	7DB3	綑		
4080	羣	12	12	82FC	7FA3	羣		
4089	脇	2	2	8346	8107	脇		
4090	脚	6	6	8347	811A	脚		
4093	舐	1	1	834A	445B		舌  氏	
4094	館	1	1	834B	8218	館		
4105	菓	2	2	8356	83D3	菓		

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
4132		25	25	8371			血△𠂇	眾，異體字問題。
4134	衛	3	3	8373	885E	衛		
4139	袴	1	1	8378	88B4	袴		
4140	衽	1	1	8379	88B5	衽		
4146	覩	6	6	83A1	89A9	覩		
4167	賁	3	3	83B6	8CEB	賁		
4174	躒	1	1	83BD	8EB6	躒		
4176	輒	8	8	83BF	8F19	輒		
4184	迹	3	3	83C7	8FF9	迹		
4195	鉢	1	1	83D2	9262	鉢		
4203	隣	1	1	83DA	96A3	隣		
4213	鞞	1	1	83E4	97BE	鞞		
4226	飡	2	2	83F1	98E1	飡		
4242	鬪	4	4	8442	9B2D	鬪		Unicode 呈現差異。
4256	鷄	31	31	8450	9DC4	鷄		
4267	鼈	4	4	845B	9F08	鼈		

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
4270	𪗇	1	1	845E	9F45	𪗇		
4283	戲	1	1	846B	622F	戲		
4307	罵	5	5	84A5	99E1	罵		
4460	勅	1	1	85A1	52D1	勅		
4507	吳	50	50	85D0	5434	吳		
4603	憇	2	2	8671	6187	憇		
4750	卅	3	3	8767			卅	
4768	妬	3	3	8779	59AC	妬		
4805	瀆	1	1	87C0	51DF	瀆		
4808	健	1	1	87C3	3493		彳 厶 達	
4814	健	3	3	87C9	5FA4	健		
4902	兗	4	4	8862	5156	兗		
5019	畧	7	7	88F9			畧	
5119	博	2	2	89C0	613D	博		
5176		10	10	89F9			彳 厶 冗	沉，異體字問題。
5290	携	22	22	8ACE	643A	携		

編號	造字	字頻 (txt)	字頻 (xml)	Big5	Unicode	WinXP	構字式	備註
5347	憾	1	1	8B48	617D	憾		
5434	惶	1	1	8BC1	6E7C	惶		
5466	倏	2	2	8BE1	5010	倏		Unicode 呈現差異。
5539	𪗇	1	1	8C6B	29C5B		𪗇𪗈	
5559	欸	8	8	8CA1	6B35	欸		
5587	欸	1	1	8CBD	6B58	欸		
5697	煑	1	1	8D6C	7151	煑		Unicode 呈現差異。
5704	羶	1	1	8D73	6C0A	羶		
5706	敍	1	1	8D75	654D	敍		

附件二、《庚巳編》手動取代表

編號	造字	檔案原文摘錄	手動取代結果
1544	𪗇	<qe>夢身中鄉試六十七名 原本闕「六」、「七」二字，據說庫本補。</qe>	<qe> 𪗇 夢身中鄉試六十七名 原本闕「六」、「七」二字，據說庫本補。</qe>
		<u> </u> (全部取代)	𪗇 (全部取代)